AI and Memory Wall

Year: 2024 | Citations: 240 | Authors: A. Gholami, Z. Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney

Abstract

The availability of unprecedented unsupervised training data, along with neural scaling laws, has resulted in an unprecedented surge in model size and compute requirements for serving/training large language models. However, the main performance bottleneck is increasingly shifting to memory bandwidth. Over the past 20 years, peak server hardware floating-point operations per second have been scaling at 3.0${\times}$× per two years, outpacing the growth of dynamic random-access memory and interconnect bandwidth, which have only scaled at 1.6 and 1.4 times every two years, respectively. This disparity has made memory, rather than compute, the primary bottleneck in AI applications, particularly in serving. Here, we analyze encoder and decoder transformer models and show how memory bandwidth can become the dominant bottleneck for decoder models. We argue for a redesign in model architecture, training, and deployment strategies to overcome this memory limitation.